REMARKS

Claim 2 has been amended to clarify the invention. In particular, claim 2 has been amended to recite the complement of SEQ ID NO:41 in the alternative. The specification has been amended to correctly reference Table 1 data. No new matter has been added by the amendments to the claims. Entry of these amendments is respectfully requested.

Applicants respectfully request that, upon allowance of product claims, method claims 4-12 and 15 should be rejoined. The Commissioner's Notice in the Official Gazette of March 26, 1996, entitled "Guidance on Treatment of Product and Process Claims in light of In re Ochiai, In re Brouwer and 35 USC § 103(b)" sets forth the rules, upon allowance of product claims, for rejoinder of process claims covering the same scope of products.

Claims Rejections - 35 USC §§ 101 and 112

1. The Examiner has rejected claims 1-3 and 13-14 under 35 USC § 101, because "the claimed invention is not supported by a specific utility because the disclosed uses of the polynucleotide are not specific and are generally applicable to any polynucleotide". The Examiner states that the specification discloses many potential uses for the polynucleotide including gene expression, as a probe, and "for expressing mRNA and protein, or conversely to block transcription or translation of the mRNA". Additionally, the Examiner states that these are "non-specific uses that are applicable to polynucleotides in general and not particular or specific to the polynucleotide claimed". Further, the Examiner states that "the claimed polynucleotide is not supported by substantial utility because no substantial utility has been established for the claimed subject matter" and "the polynucleotide is known only to be co-expressed with creatine kinase M, and has no asserted or identified specific and substantial utilities". Furthermore, the Examiner states that, the claimed invention is "generic in nature and applicable to a myriad of such compounds" and "Neither the specification as filed nor any art of record discloses or suggests any property or activity for the polynucleotides such that another non-asserted utility would be well established for the compounds. This rejection is traversed.

Applicants hereby submit DECLARATION UNDER 37 CFR 1.132 OF MICHAEL G. WALKER. As shown in the Declaration and now well known in the art, the claimed combination is diagnostic of disorders of cardiac muscle function. In the Declaration, Dr. Walker discusses that by the Guilt-by-Association, or GBA method, out of a total of *45,233* assembled sequences, only *48* novel polynucleotides showed strong expression and association with the known cardiac muscle-associated genes. Contrary to the Examiner's statement, the asserted utility is not applicable to all polynucleotides

since 45,185 sequences (45,233 minus 48) were not significantly expressed in or associated with disorders of cardiac muscle function. This specific utility was asserted at the time of filing and is described in the specification under the invention section on pages 5-8, and in particular, on page 6, line 2, and on page 48, lines 7-8. Thus, the Declaration presents evidence from an inventor who is also a person skilled in the art to support Applicants' asserted utility for the claimed invention as surrogate markers for disorders of cardiac muscle function, and specifically for cardiac injury, which were identified by their co-expression with known diagnostic and prognostic markers.

In further support, Dr. Walker presented a recent article by Thompson et al. (2002, "Identification and Confirmation of a Module of Coexpressed Genes", Genome Research 12:1517-1522), which validates the GBA method for predicting coexpression of genes (see p. 1518 of Thompson et al.). In particular, as presented in the DECLARATION, the Thompson article establishes that a set of genes coexpressed in breast and ovarian cancer as determined by GBA analysis were indeed coexpressed as determined by real-time, RT-PCR analysis of mRNA extracted from cell lines derived from breast tissue (see p. 1520 and Figure 2 of Thompson et al.). It is further noted that the Thompson study used the method of analysis of the instant application applied to data from the GenBank EST database, data which is also included in the LIFESEQ database. Therefore, based on the arguments presented above, the teachings in the specification, the Thompson article and the DECLARATION UNDER 37 CFR 1.132 OF MICHAEL G. WALKER, Applicants submit that one skilled in the art would more likely than not concur with Applicant's asserted characterization that the claimed combination is diagnostic of disorders of cardiac muscle function and has specific and substantial utility.

Applicants respectfully point to the information supplied in the DECLARATION UNDER 37 CFR 1.132 OF MICHAEL G. WALKER which clearly demonstrates that SEQ ID NO:41 was selected as a representative for the combination and as a surrogate marker for the very highly associated creatine kinase M, a known diagnostic marker for cardiac injury. The utility for the combination, SEQ ID NO:41, or an antibody that specifically binds the gene product of SEQ ID NO:41 is diagnosis of disorders of cardiac muscle function as defined in the specification on page 4, lines 15-18, of the specification. The DECLARATION also states that the prognosis of cardiac patients is greatly improved with early diagnosis and treatment. Therefore, the present specification identified unknown genes that can be used as surrogate markers in that diagnostic process without even knowing a priori the name, structure, or function of the gene or the encoded protein.

By virtue of this DECLARATION, Applicants have more than exceeded the "practically useful", and "specific benefit" to the public standards for 35 USC § 101 utility as described in Anderson v. Natta,

(480 F2d 1392, 1397, 178 USPQ 458; CCPA 1973) and Brenner v. Manson (383 US 519, 534-35, 148 USPQ 689; 1966). Clearly, an invention that can be used to diagnose lung cancer is both practically useful and confers a specific benefit to the public. Applicants respectfully point out that the threshold for demonstration of utility is "not statistical certainty" as promulgated by the court in Nelson v. Bowler (626 F.2d 853, 856-57, 206 USPQ 881, 883-84). Furthermore, and in contrast to criminal cases, the applicant does not have to provide evidence sufficient to establish that an asserted utility is true "beyond a reasonable doubt" (In re Irons, 340 F.2d 974, 978, 144 USPQ 351, 354; CCPA 1965; MPEP 2107.02, section VII; CCPA 1980).

Applicants respectfully submit by virtue of the disclosure in the specification as filed and with the Declaration of Dr. Walker who is clearly a person skilled in the art, that the claimed invention is more than capable of providing an identifiable benefit--the claimed polynucleotides for the diagnostic of disorders of cardiac muscle function.

With the DECLARATION UNDER 37 CFR 1.132 OF MICHAEL G. WALKER, the Thompson article, the arguments, and explanations above, Applicants respectfully request that the rejection of claims 1-3, 13 and 14 under 35 USC § 101 be withdrawn.

2. The Examiner has rejected claims 1-3 and 13-14 under 35 USC § 112, first paragraph, based on the rejection of these claims for lack of utility under 35 USC § 101.

Applicants respectfully submit that they have provided sufficient evidence to clearly demonstrate that the 35 USC § 101 utility requirement has been satisfied. Therefore, Applicants respectfully request withdrawal of the rejection of claims 1-3 and 13-14 under 35 USC § 112, first paragraph.

CONCLUSION

In light of the above amendments and remarks, Applicants submit that the present application is fully in condition for allowance and request that the Examiner withdraw the outstanding rejections. Early notice to that effect is earnestly solicited.

If the Examiner contemplates other action or if a telephone conference would expedite allowance of the claims, Applicants invite the Examiner to contact Applicants' Agent at (650) 855-0555.

Applicants believe that no fee is due with this communication. However, if the USPTO determines that a fee is due, the Commissioner is hereby authorized to charge Deposit Account No. **09-0108.**

Respectfully submitted,
INCYTE GENOMICS, INC.

Date: 13 November 2002

Yu-Mei Eureka Wang
Reg. No. 50,510
Direct Dial Telephone: (650) 621-8740

3160 Porter Drive
Palo Alto, California 94304
Phone: (650) 855-0555
Fax: (650) 849-8886

VERSION WITH MARKINGS TO SHOW CHANGES MADE

IN THE SPECIFICATION

Paragraph(s) beginning at line 20 of page 23 has been amended as follows:

The data in Table 1-1, 1-2, and 1-3 [above] can be summarized by reducing it to a single highest co-expression (-log p) value for each intersecting known gene and unknown polynucleotide and naming at least one disorder associated with expression of the known gene. A summary table is shown below:

IN THE CLAIMS

Claim 2 has been amended as follows:

2. (Twice Amended) An isolated polynucleotide comprising a nucleic acid sequence of SEQ ID NO:41 or [and] the complement of SEQ ID NO:41.

FB-0009-1CIP

CERTIFICATE OF TRANSMISSION
I hereby certify that this paper is being facsimile transmitted to the attention of Examiner David R. Gunter, Group Art Unit 1634, USPTO, to Facsimile No. 703-746-9212 on 13 November 2002.

Signature of Margaret Nielsen

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of: Walker et al.
Title: **POLYNUCLEOTIDES ASSOCIATED WITH CARDIAC MUSCLE FUNCTION**

Serial No.:	09/880,192	Filing Date:	June 12, 2001
Examiner:	Gunter, D.	Group Art Unit:	1634

Commissioner for Patents
Washington, DC 20231

DECLARATION UNDER 37 CFR 1.132 OF MICHAEL G. WALKER

I, Michael G. Walker, declare:

1. I received my doctoral degree from Stanford University, Stanford, CA in 1992. I have held the position of Consulting Professor in the Department of Medicine at Stanford University since 1995 and have consulted to Incyte Genomics since 1996. My work has involved the development of analytical tools for the characterization and annotation of molecules by their expression in relation to cellular function, disease, and metabolic pathways. I am the first named inventor on the pending application.

2. The application relates to polynucleotides that are surrogate markers for disorders of cardiac muscle function, and specifically for cardiac injury, which were identified by their co-expression with known diagnostic and prognostic markers.

3. I understand that the OFFICE ACTION presents 35 USC §§ 101 and 112 rejections of claims 1-3, 13, and 14 directed to the described invention as "not supported by a specific utility because the disclosed uses of the polynucleotide are not specific and generally applicable to any polynucleotide".

4. The purpose of my declaration is to support the asserted utility of the identified polynucleotides as diagnostics. To that end, I will discuss the attached journal article of Thompson et al. (2002; Identification and Confirmation of a Module of Coexpressed Genes, Genomics Research 12:1517-1522) and the data presented in the application.

5. I will now summarize the relevant points of the Thompson et al. article.

a) Thompson et al. used guilt-by-association method (GBA, the method used in this application)

across all tissues of the dbEST and UniGene databases to identify functional modules of gene expression data (all genes similarly expressed across all tissues).

b) The data set was tested and found to be reliable using ubiquitously expressed genes and known coexpressed genes.

c) A novel functional module was found to be involved in breast cancers.

d) Quantitative expression profiles of six of the genes identified in the novel functional module were confirmed using reverse transcriptase real-time PCR with four different cell lines derived from mammary epithelial tissue.

e) Published information on these six genes revealed interactions in two distinct classes: transcriptional control and ubiquitin proteasome pathway. Thompson et al. suggest that three of the genes in the first class regulate transcription of the three genes in the second class.

f) Thompson et al. stated that their work focused on a small subset of the available data and is expected to lead to the identification of many more functional modules.

7. I will now point out the parts of the specification that support the asserted utility, that the claimed combination is diagnostic of disorders of cardiac muscle function.

a) The known genes used in GBA were characterized as to function and disease indication in EXAMPLE V.

b) As recited in EXAMPLE VI, "GBA identified 48 novel polynucleotides from a total of 45,233 assembled sequences that showed strong expression and association with the known cardiac muscle-associated genes". Contrary to the Examiner's statement, the asserted utility is not applicable to all polynucleotides since 45,185 sequences were not significantly expressed in or associated with disorders of cardiac muscle function.

c) The known and unknown genes, the p-value for their highest association, and their disease indication were summarized in a table at the end of EXAMPLE VI. It is clear from the table that cardiac injury is the broadest diagnosis using the claimed combination of polynucleotides and that specific problems such as atrial fibrillation and myocardial infarction are further identifiable using the expression of their associated known and unknown markers.

d) SEQ ID NO:41 was selected as a representative for the combination and as a surrogate marker for the very highly associated creatine kinase M, a known diagnostic marker for cardiac injury. Therefore the utility for the combination, SEQ ID NO:41, or an antibody that specifically binds the gene product of SEQ ID NO:41 is diagnosis of disorders of cardiac muscle function as defined in the

specification on page 4, lines 15-18, of the specification.

8. Finally, it is now well known that the prognosis of cardiac patients is greatly improved with early diagnosis and treatment. The present specification identified unknown genes that can be used as surrogate markers in that diagnostic process without even knowing a priori the name, structure, or function of the gene or the encoded protein.

I hereby declare that all statements made herein are true and that they are based on my own knowledge, information and belief. These statements are made with the knowledge that willful false statements are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of this application or any patent issued from it.

Date: 12 Nov 02

Michael G. Walker, PhD

Incyte Genomics
3160 Porter Drive
Palo Alto CA 94304

Letter

Identification and Confirmation of a Module of Coexpressed Genes

H. Garrett R. Thompson,[3] Joseph W. Harris,[3] Barbara J. Wold,[1] Stephen R. Quake,[2] and James P. Brody[3,4]

Departments of [1]Biology and [2]Applied Physics, California Institute of Technology, Pasadena, California 91125, USA; [3]Department of Biomedical Engineering, University of California Irvine, Irvine, California 92697, USA

We synthesize a large gene expression data set using dbEST and UniGene. We use guilt-by-association (GBA) to analyze this data set and identify coexpressed genes. One module, or group of genes, was found to be coexpressed mainly in tissue extracted from breast and ovarian cancers, but also found in tissue from lung cancers, brain cancers, and bone marrow. This module contains at least six members that are believed to be involved in either transcritional regulation (PDEF, H2AFO, NUCKS) or the ubiquitin proteasome pathway (PSMD7, SQSTM1, FLJ10111). We confirm these observations of coexpression by real-time RT–PCR analysis of mRNA extracted from four model breast epithelial cell lines.

Molecular studies of cellular functions have led to broad knowledge of cellular processes. Most cellular processes are the result of molecules interacting, rather than due to the activity of individual molecules. The study of functional groups of genes has been termed modular cell biology (Hartwell et al. 1999). We are interested in identifying functional modules.

Technical constraints make the study of individual genes much easier than identifying interacting molecules. One approach to identify modules is to examine genome scale expression data.

Genes coexpressed in many different tissues, under both normal and diseased conditions, and at different times during development, are candidates for forming functional modules.

There are different types of experimental gene expression data sets available for use to identify modules. DNA microarray experiments measure mRNA expression levels on a known set of genes by way of hybridization. A fluorescent tag is put on an unknown single-stranded DNA molecule. A set of single-stranded DNAs of known sequence is immobilized onto a surface at known locations. The unknown fluorescently tagged sample is allowed to hybridize to its complimentary immobilized strand. The surface is scanned to create a fluorescent image. The intensity of the fluorescence is a measure of the concentration of DNA in the sample.

Substantial DNA microarray data sets exist, but are not easy to compare between different laboratories. These measurements are made relative to a control and the numbers are reported as a fold difference relative to the control. The specific choice for a control varies widely between different laboratories. Most troubling is the lack of any reported error in these measurements. Hence, there is little systematic information available on how certain the experimenter is of the reported value.

Similar expression data measurements can be made by DNA sequencing. Message RNA molecules isolated from a tissue are reverse transcribed into cDNA and cloned into *Escherichia coli* vectors to generate a library. Random clones are sampled from the library and a few hundred base pairs are sequenced from each. These are known as expressed sequence tags (ESTs). The sequence read from each clone is generally sufficient to identify the gene when cross referenced to a consensus sequence data set.

The UniGene data set (Schuler et al. 1996) groups the large number of publicly available DNA sequence fragments on the basis of sequence overlaps into unique genes. In addition to complete sequences of some well-known genes, there are thousands of uncharacterized EST fragments that are found as a result of the clustering process. These uncharacterized EST fragments are thought to represent previously uncharacterized genes.

Recently, concerted efforts have been formed to construct a diverse set of cDNA libraries derived from various tissues in normal and diseased states. These libraries are heavily influenced by the National Cancer Institute's Cancer Genome Anatomy Project (CGAP), whose stated goals are to characterize normal, precancerous, and malignant cells. These libraries, when combined with the UniGene collection, provide a set of gene expression data that can be analyzed to identify groups of coexpressed genes.

In general, EST sequencing is more accurate but substantially more expensive than DNA microarrays. DNA microarrays suffer from cross-hybridization, uncertain linearity, and numerous other technical problems. However, they are almost always more sensitive and cheaper to use. (The sensitivity of EST sequencing is limited by the depth of sampling into the library. The costs increase linearly with the depth of sequencing.) However, EST sequencing provides data that is easily comparable across different experiments. EST sequencing provides absolute numbers that represent a sampling of the mRNAs present in a tissue sample.

The specific purpose of this analysis is to identify functional modules from gene expression data by identifying groups of genes that are coexpressed. Our hypothesis is that coexpression has functional significance. Our approach of analyzing gene expression across all tissues (rather than just comparing pairs of diseased and normal tissue), offers two advantages. First, because only genes similarly expressed across all tissues will be grouped together, the identification of both tissue-specific and tissue-generic modules is possible.

[4]Corresponding author.
E-MAIL jpbrody@uci.edu; FAX (949) 824-9968.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.418402.

Table 1. The Ten Most Ubiquitously Expressed Genes in the Dataset

Total	Gene	Description
422	EEF1A1	eukaryotic translation elongation factor 1 alpha 1
422	ACTB	actin, beta
371	ACTG1	actin, gamma 1
356	GAPD	glyceraldehyde-3-phosphate dehydrogenase
354	RPLP0	ribosomal protein, large, P0
353	EEF1G	eukaryotic translation elongation factor 1 gamma
342	TPT1	tumor protein, translationally-controlled 1
334	RPL13A	ribosomal protein L13a
330	HDAC3	histone deacetylase 3
326	RPS4X	ribosomal protein S4, X-linked

There were 1573 libraries examined and the above ran appeared in the greatest number of libraries. Total indicates the number of libraries (of 1573) in which that particular gene was found to be expressed.

Second, modules of genes that have a specific and normal function in, for instance, the development of the fetal brain, but also play a pathologic role in, for example, some forms of cancer due to mutations that lead to misregulation may be identified.

Although dbEST (Benson et al. 2002) contains sequence data from many different laboratories, the vast majority of sequence analyzed in this study is derived either from the Washington University/Merck EST project (Hillier et al. 1996) or the National Cancer Institute's Cancer Genome Anatomy Project (CGAP).

The simple listing of ESTs, as is provided in dbEST, is insufficient for large-scale gene expression analysis. A clustering (or assembly) of the ESTs into unique genes is also needed. At least two databases have been established to do this. The first, UniGene (Wheeler et al. 2002), is regularly updated and available freely online. The second, TIGR Human Gene Index (Quakenbush et al. 2000), is available with a restrictive license. The data presented in this study is based on the analysis of gene expression data compiled in the UniGene database.

Analysis algorithms for gene expression data fall into two classes. The first identifies differentially expressed genes between two experiments (cells under different conditions, tissues, or groups of tissues). This is most simply done with a Fisher's Exact test, but can also use more rigorous statistical methods (Audic and Claverie 1997; Stekel et al. 2000). The second class of algorithms seeks to identify groups of genes with similar patterns of expression across many (hundreds) different experiments. We need an algorithm from this second class.

Many different approaches have been used to classify and organize gene expression data. Most algorithms are variations on Eisen's approach (Eisen et al. 1998) of organizing the data on the basis of similarity of gene expression. A different approach, guilt-by-association (GBA) was described by Walker et. al (1999) and used to identify novel genes with expression patterns similar to those of known disease-related genes in a private (Incyte's LifeSeq) gene expression data set generated from ESTs. In this work, we used the GBA algorithm to identify genes with similar expression patterns.

In this study, we first constructed a large set of human expression data from dbEST and UniGene. The data set was found to be reliable on the basis of its identification of ubiquitously expressed genes and known coexpressed genes. Then, we analyzed this data set to identify modules of coexpressed genes. The analysis identified potential functional modules. One was identified in the literature as a module expressed during pregnancy (Thompson et al. 1990), further confirming the reliability of the data set. We also identified a set of coexpressed genes that may form a novel functional module. The members of this functional module are not well studied in the literature. We examined gene expression in several cell lines to confirm that these genes are coexpressed.

RESULTS

After constructing the gene expression data set, we first examined ubiquitously expressed genes as a check on the validity of the data set. Genes that showed the widest expression (found in the greatest number of libraries) are shown in Table 1. No genes appear in more than one-third of the libraries.

Table 2. A Module of Genes That Is Expressed During Pregnancy

Coinc	Total	Gene	Description
9	16	PSG4	pregnancy specific β-1-glycoprotein 4
10	65	CHS1	chorionic somatomammotropin hormone 1 (placental lactogen)
11	23	PSG9	pregnancy specific β-1-glycoprotein 9
7	8	PSG3	pregnancy specific β-1-glycoprotein 3
9	12	PSG6	pregnancy specific β-1-glycoprotein 6
7	13	PSG2	pregnancy specific β-1-glycoprotein 2
7	22	PAPPA	pregnancy-associated plasma protein A
7	15	CSH2	chorionic somatomammotropin hormone 2
10	17	CYP19	cytochrome P450, subfamily XIX (aromatization of androgens)
10	30	DAM12	a disintegrin and metalloproteinase domain 12 (meltrin alpha)
8	15	PSG5	pregnancy specific β-1-glycoprotein 5
11	46	TFPI2	tissue factor pathway inhibitor 2
5	5	PSG7	pregnancy specific β-1-glycoprotein 7
11	356	GAPD	glyceraldehyde-3-phosphate dehydrogenase

The genes are listed relative to PSG1 (pregnancy specific β-1-glycoprotein), which is found in 13 different libraries. The table lists (coinc) the number of libraries in which both PSG1 and the listed gene are both found and (total) the number of libraries in which the listed gene is found, but PSG1 is not found. For example, PSG4 is found in a total of 16 libraries, PSG1 is present in a total of 13 libraries. In 9 of the 13 libraries that PSG1 is found in, PSG9 is also found. GAPD is shown for reference.

Table 3. A Module (the PDEF Module) of Genes That Is Expressed in Breast and Ovary Cancers

Coinc	Total	Gene	Description
9	46	H2AFO	H2A histone family, member O
22	105	PSMD7	proteasome (prosome, macropain) 26S subunit, non-ATPase, 7 (Mov34 homolog)
16	104	NUCKS	similar to rat nuclear ubiquitous casein kinase 2
14	51	FLJ10111	hypothetical protein FLJ10111
21	203	SQSTM1	sequestosome 1
10	356	GAPD	glyceraldehyde-3-phosphate dehydrogenase

The genes are listed relative to PDEF (prostate derived Ets factor, a transcription factor), which is found in 27 different libraries. The table lists (Coinc) the number of libraries in which both PDEF and the listed gene are both found and (Total) the number of libraries in which the listed gene is found, but PDEF is not found. GAPD is shown for reference.

The results of our analysis for coexpressed genes revealed both known relationships and unknown relationships. We highlight one of the known relationships as anecdotal evidence for the effectiveness of this approach. Table 2 shows the 10 genes with expression patterns most similar to pregnancy specific β-*1-glycoprotein*. The pregnancy-specific glycoproteins are a group of proteins that are found in large amounts in placenta (Thompson et al. 1990). They are all located on chromosome 19 and share common transcriptional regulatory elements (Thompson et al. 1990).

A postulated functional module is shown in Table 3. These five genes are all shown with their coexpression relative to *PDEF*, a recently described transcription factor (Oettgen et al. 2000). To better understand the role of this module, we examined the tissues in which it was expressed. We found the module to be expressed in some breast (Table 4) and ovary (Table 5) tissues. The module is also expressed in several other brain, lung cancer, and bone marrow tissues (Table 6).

Many other significant relationships were found between pairs of genes. Significance of a relationship was defined, following Walker et al. (1999), by a P-value, indicating the probability that the coexpression was due to chance. Over 29,000 gene pairs showed a P-value of $<10^{-6}$. The distribution of these P-values is shown in Figure 1 along with the approximate range of P-values for the functional modules described in this work. There is not necessarily a clear relationship of what constitutes a module. The structure of the data is more like a complex web then a group of distinct modules and our definition of which genes belong to the PDEF module is probably not complete.

To confirm the observation of coexpression, we performed quantitative reverse transcriptase real-time PCR on four different cell lines derived from mammary epithelial tissue. We measured the quantity of mRNA from *PDEF* and the five other genes shown in Table 3, along with a control, *GAPD*. Results are shown in Figure 2.

DISCUSSION

Examination of the most ubiquitously expressed genes (Table 1) provides interesting insight into the typical population of cellular mRNAs. As expected, most of these mRNAs code for well-known ubiquitously expressed proteins (components of the ribosome, structural proteins, and enzymes). However, one protein labeled tumor protein, translationally controlled 1 or TPT1, is prominent due to its obscurity in the literature. Although its name implies association with cancer, it is better described as a histamine releasing factor (HRF) and has been studied as playing a role in the allergic response of mast cells that results in the exocytosis of histamine. However, its ubiquitous nature perhaps indicates it plays a much larger role in cellular signaling. In fact, the recent determination of the structure of this protein (Thaw et al. 2001) revealed significant similarity to Mss4, a small GTPase accessory protein involved in amplification of extracellular signals. Our finding that TPT1 is ubiquitously expressed is supported by other works.

Table 4. The Expression of the PDEF Module in Libraries Derived from Breast Tissues

ID	Description
517	invasive ductal breast tumors (pooled bulk)
557	ductal tumor (bulk)
590	infiltrating ductal carcinoma (microdissected)
634	normal (bulk)
730	• carcinoma: in situ (microdissected)
759	• carcinoma: invasive (microdissected)
766	• adenocarcinoma (microdissected)
768	• carcinoma: lobular (microdissected)
770	• adenocarcinoma (microdissected)
931	high grade neoplasia (bulk)
726	• normal epithelium (microdissected)
583	normal ductal tissue (microdissected)

The bullets indicate in which libraries the PDEF module is present. ID is the UniGene library identification number.

Table 5. The Expression of the PDEF Module in Libraries Derived from Ovary Tissues

ID	Description
186	tumor/normal (bulk)
389	normal (bulk)
390	71 yrs old normal (bulk)
514	invasive tumor serous, papillary, adenocarcinoma (micro-bulk)
564	serous papillary adenocarcinoma (bulk)
652	serous adenocarcinoma (bulk)
675	• serous papillary carcinoma (microdissected)
706	• borderline neoplasia (microdissected)
708	• borderline preneoplasia (microdissected)
709	• invasive carcinoma (microdissected)
710	• invasive carcinoma (microdissected)
731	• borderline preneoplasia (microdissected)
765	serous papillary carcinoma (microdissected)
717	serous papillary, clear cell, spindle cell, carcinoma (bulk)

The labels are identical to those in Table 2.

Table 6. The PDEF Module Is Also Expressed in Libraries Derived from Other Tissues

UniGene ID	Description
764	• adult bone marrow normal stem cells 34+/38+ (flow sorted)
767	• adult brain oligodendroma (bulk)
771	• adult lung carcinoma: bronchioloalveolar (microdissected)
774	• adult lung adenocarcinoma: invasive (microdissected)
812	• adult bone marrow (lymphoid tissue) normal stem cell (bulk)
819	• adult brain oligodendroglioma (bulk)
857	• fetus brain normal (bulk)

The labels are identical to those in Table 2.

One study used SAGE to exhaustively enumerate the populations of mRNAs within colorectal cancer cell lines (Velculescu et al. 1999) and found TPT1 to be among the most abundant mRNAs in those cell lines. A second study found a yeast homolog to TPT1 to be one of the top 20 most abundant proteins by two-dimensional gel analysis (Norbeck and Blomberg 1997).

The reliability of the data set was further confirmed by examining genes coexpressed with *PSG1*. The human pregnancy-specific glycoproteins (PSGs) form a group of proteins that are closely clustered on chromosome 19. This group of genes, along with a few others (*CSH1*, *CSH2*) were found in libraries derived from placenta and fetal tissues. Interestingly, one library derived from aorta (UniGene Library ID 182) also contained many of these genes. All of these genes are known to be specifically expressed in placental tissues.

We also identified members of a novel functional module. Six members of this postulated functional module are reported as follows: PDEF, SQSTM1, FLJ10111, H2AFO, PSMD7, and NUCKS.

PDEF was first identified as a prostate epithelium-specific Ets transcription factor that interacts with the androgen receptor to activate the promoter of the well-known prostate cancer marker gene, *PSA* (Oettgen et al. 2000). Simultaneously, it was identified and cloned in both human (Nozawa et al. 2000) and mouse (Yamada et al. 2000), in which it was identified as a positive regulator of maspin, a protease inhibitor that is down-regulated in advanced breast cancers (Zou et al. 1994). Later investigators identified PDEF as a candidate breast tumor marker (Ghadersohi and Sood 2001) and showed that PDEF mRNA was highly overexpressed in 14 of 20 primary human breast tumors examined. They also found that one patient with metastatic breast cancer had PDEF mRNA 192-fold higher in blood as compared with normal individuals (Ghadersohi and Sood 2001). Although PDEF mRNA has been identified as being highly up-regulated in breast cancer cell lines (Fig. 1) and breast tumors (Ghadersohi and Sood 2001), only two targets [the PSA promoter (Oettgen et al. 2000) and the maspin promoter (Yamada et al. 2000)] have been identified that PDEF regulates. Little is known about PDEF's function, either in breast cancer or in normal tissue.

The class of Ets transcription factors is characterized by the ETS domain, which binds monomerically to an 8-bp long DNA element (Dittmer and Nordheim 1998). Most Ets transcription factors bind to the GGAA core sequence of DNA, but PDEF prefers a GGAT core (Oettgen et al. 2000).

P62 is a widely expressed cytoplasmic protein that is known to play a role in cellular signaling by interacting with the SH2 domain of p56(lck) (Park et al. 1995). P62 was also identified as a ubiquitin-binding protein through a yeast two-hybrid screen (Vadlamudi et al. 1996). More recently, p62 has been found to form cytoplasmic structures called intrahyaline bodies that serve as a storage place for multiubiquitinated proteins (Shin 1998). The ability of p62 to bind noncovalently to ubiquitin and several signaling proteins, suggests that p62 may play a regulatory role connected to the ubiquitin–proteasome system.

The promoter of the *SQSTM1* gene has been characterized previously (Vadlamudi and Shin 1998). This information, along with the DNA-binding studies of Oettgen et. al (2000), allowed us to identify two potential PDEF-binding sites within the *SQSTM1* promoter and to postulate that PDEF regulates the *SQSTM1* promoter. We have performed transfection experiments showing that PDEF activates the *SQSTM* promoter (Thompson et al. 2002) threefold over basal levels.

The hypothetical protein FLJ10111 was first identified and named by the Nedo Full-length cDNA Sequencing Project in Japan. Although the message for the gene has been found

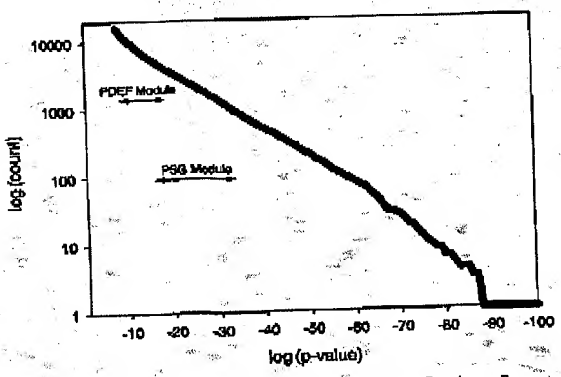

Figure 1 The cumulative distribution of pairwise P-values. For example, ~1000 gene pairs showed correlations with a P-value $<10^{-30}$ and ~100 showed correlations with a P-value $<10^{-40}$. The P-values for each gene pair in the modules described in this work fall within the indicated regions. Many other significant coexpressions were observed, but not confirmed.

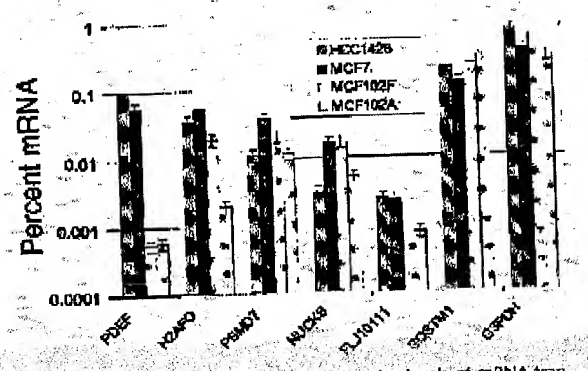

Figure 2 Real-time PCR data quantifying the levels of mRNA transcript for each indicated gene in four different cell lines.

(Yawata et al. 2001), no gene product has yet been characterized. Several lines of evidence point to the gene product being involved in the ubiquitin–proteasome pathway.

FLJ10111 is located in a tightly packed region of chromosome 14 that contains other proteasome-related genes (Yawata et al., 2001). This 35-kb region contains six different genes, two of which (*PA28α* and *PA28β*) code for components of the proteasome activator (PA28), and transcription of genes in this region is induced by γ-interferon – suggesting a mode of regulation. The PA28 proteasome activator is thought to bind to the 20 S proteasome and enhance the generation of major histocompatibility complex (MHC) class I-binding peptides (Ma et al. 1992; Yawata et al 2001). The products of the other four genes in this 35-kb region are not well characterized. The predicted protein encoded by FLJ10111 has two RING finger domains, which are characteristic of E3 ubiquitin ligases (Yawata et al. 2001). The RING finger motif is found in many (>200) proteins in different eukaryotes, but not in any prokaryote proteins (Freemont 2000).

Histones are nuclear proteins responsible for organizing genomic DNA into the tightly packed chromosomes within eukaryotic cells. The histones play a role in eukaryotic transcriptional regulation and post-translational modifications can change their DNA-binding properties. The transcript for H2AFO, a member of the H2A family, was found among this group of genes.

The protein encoded by the *PSMD7* gene, S12, is a regulatory subunit of the proteasome. This protein is homologous to the mouse Mov34 protein. Mutations in Mov34 are lethal in the embryonic stage of development. The protein is evolutionary conserved and homologs can be found in yeast, *Drosophila*, along with the mouse.

NUCKS is similar (by sequence analysis) to the rat nuclear ubiquitous casein kinase 2, which was first isolated from both HeLa cells and rat brain and subsequently found to be ubiquitously expressed (Ostvold et al. 2001). The protein localizes to the nucleus and binds single- and double-stranded DNA. It is phosphorylated at multiple sites by several kinases, including protein kinase C and casein kinase 2 (Ostvold et al. 1985, 1992; Walaas et al. 1989).

This approach to identifying functionally related modules of genes has wide applicability. The libraries that comprise the data set we examined were largely populated by those sequenced under the support of the NCI's CGAP program, which is focused on cancer. Hence, it is not surprising that we identified a cancer-related module. We expect that concerted expression profiling of other diseased tissues will uncover functional modules active in those diseases. This method, however, requires a discovery-based approach to studying molecular interactions. We queried the data set to examine some well-known genes and there is no information available on these. This may be because these genes have rarely (rather than abundantly) expressed mRNAs and the paucity of sequences that were read in each library.

The hypothetical functional module that we identified in this study has members in two distinct classes, transcriptional control (H2AFO, PDEF, and NUCKS) and ubiquitin–proteasome pathway (PSMD7, *SQSTM1*, and FLJ10111). The obvious hypothesis that this leads to is that members of the first class control the transcription of the second class and that this occurs in some diseased tissues. We have shown that PDEF activates the SQSTM1 promoter and that the product of *SQSTM1*, p62, is overexpressed in breast cancer samples relative to normal breast tissue (Thompson et al. 2002).

In summary, we use a method to identify a hypothetical functional module of coexpressed genes in publicly available large-scale gene expression data. We confirmed that these genes are coexpressed by quantitative measurements of mRNA in cell lines derived from breast tissue. Published information on the function of the products of these genes leads to a hypothesis about how these are functionally related. This work focuses on a small subset of the available data. The approach has wide applicability and could lead to the identification of many more functional modules.

METHODS

Computation

We scanned the Homo Sapiens UniGene data set (build #146, available at ftp://ftp.ncbi.nlm.nih.gov/repository/UniGene/) and compiled a list of libraries in which each UniGene appears. Some appear multiple times in a single library. We screened out all libraries that had less than 100 UniGenes associated with it. The final data set that we analyzed consisted of expression information for 96,574 genes across 1573 libraries.

We analyzed gene expression data by using guilt-by-association (GBA), a combinatoric measure of similarity between the expression patterns of two genes (Walker et al. 1999). We calculated the pairwise expression similarity between all 96,574 genes (4.7 billion comparisons).

Cell Culture

Four mammary epithelial cell lines were obtained from ATCC and examined for mRNA levels. MCF-7 cells were derived originally from an adenocarcinoma. These cells are estrogen-receptor positive. HCC1428 cells were first derived from a metastatic adenocarcinoma. The patient had a family history of breast cancer. The cells are Her2/neu negative and p53 negative. MCF-10-2F and MCF-10-2A cells are both nontumorigenic cell lines (they do not form tumors in immunosuppressed mice) that were derived from a 36-year-old female with fibrocystic breast disease. The MCF-10-2F cells are derived from floating cells, whereas the MCF-10-2A cells are derived from adherent cells.

Mammary epithelial cells were cultured in monolayers in 10-cm^2 culture dishes in the recommended medium supplemented with chelated or unchelated horse or fetal bovine serum as applicable, until <80% confluent. Ten dishes of cultured cells, or ~5–10 × 10^7 cells, were solubilized in TRI Reagent per the manufacturer's instructions. RNA extraction, precipitation, and solubilization were performed as described by the manufacturer. A total of 0.25 mg RNA was aliquoted for recovery of poly(A) RNA from each cell line cultured using a spin-column format (QIAGEN) per the manufacturer's instructions. Poly(A) RNA was quantitated by fluorescence using the RiboGreen RNA Quantitation Reagent (Molecular Probes) in 96-well plate format, assayed in triplicate, using the Fluoroskan Accent FL combination luminometer/fluorometer (LabSystems). A total of 50 ng mRNA was used for RT–PCR using the Thermoscript RT–PCR System (GIBCO-BRL/Invitrogen) to generate cDNA per the manufacturer's instructions.

Real-Time PCR

Prior to assaying concentrations of mRNA species from samples by real-time quantitative PCR, several preparatory steps were required. First, separate standard curves were generated for each gene to be analyzed. Genes to be analyzed were amplified from an appropriate cDNA source (i.e., prostate cDNA from Clontech) by PCR using primers designed from gene sequences available at www.ncbi.nlm.nih.gov. Amplified sequences were gel purified and quantitated by spec-

troscopy in triplicate. Tenfold serial dilutions of the standard curve over five orders of magnitude were generated and assayed in duplicate with each sample.

Primers used in the real-time PCR reaction were designed from sequences internal to those used to amplify the cDNA for the standard curve, with care taken to encompass all mRNA variants reported, in case multiple cDNAs were generated by RT–PCR. In addition, real-time primers were designed to come from two different exons to eliminate the possibility of genomic DNA and immature RNA contamination. Results from real-time PCR are collected as fluorescence over cycle number. By establishing a fluorescence threshold, a linear graph is generated of the standard curve using log (concentration) plotted as a function of cycle threshold number. The equation of the line is then used to determine the starting concentration of mRNA for each unknown. For each assay, unknowns are assayed in triplicate with a new standard curve. Samples are normalized to total mRNA but a housekeeping gene, *GAPD*, was also quantitated.

ACKNOWLEDGMENTS

This work was supported by the National Human Genome Research Institute through grant 5K22-HG000047.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Audic, S. and Claverie, J.M. 1997. The significance of digital gene expression profiles. *Genome Res.* **7:** 986–995.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A., and Wheeler, D.L. 2002. Genbank. *Nucleic Acids Res.* **30:** 17–20.

Dittmer, J. and Nordheim, A. 1998. Ets transcription factors and human disease. *Biochim. Biophys. Acta.* **1377:** F1–F11.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.* **95:** 14863–14868.

Freemont, P.S. 2000. Ring for destruction. *Curr. Biol.* **10:** R84.

Ghadersohi, A. and Sood, A.K. 2001. Prostate epithelium-derived Ets transcription factor mRNA is overexpressed in human breast tumors and is a candidate breast tumor marker and a breast tumor antigen. *Clin. Cancer Res.* **7:** 2731.

Hartwell, L.H., Hopfield, J.J., Leibler, S., and Murray, A.W. 1999. From molecular to modular cell biology. *Nature* **402:** C47–C52.

Hillier, L., Lennon, G., Becker, M., Bonaldo, M., Chiapelli, B., Chissoe, S., Dietrich, N., Dubuque, T., Favello, A., Gish, W., et al. 1996. Generation and analysis of 280,000 human expressed sequence tags. *Genome Res.* **6:** 807–828.

Ma, C.P., Slaughter, C.A., and DeMartino, G.N. 1992. Identification, purification, and characterization of a protein activator (PA28) of the 20 S proteasome (macropain). *J. Biol. Chem.* **267:** 10515–10523.

Norbeck, J. and Blomberg, A. 1997. Two-dimensional electrophoretic separation of yeast proteins using a non-linear wide range (pH 3–10) immobilized pH gradient in the first dimension; reproducibility and evidence for isoelectric focusing of alkaline (pI >7) proteins. *Yeast* **13:** 1519–1534.

Nozawa, M., Yomogida, K., Kanno, N., Nonomura, N., Miki, T., Okuyama, A., Nishimune, Y., and Nozaki, M. 2000. Prostate-specific transcription factor hPSE is translated only in prostate epithelial cells. *Cancer Res.* **60:** 1348–1352.

Oettgen, P., Finger, E., Sun, Z., Akbarali, Y., Thamrongsak, U., Boltax, J., Grall, F., Dube, A., Weiss, A., Brown, L., et al. 2000. PDEF, a novel prostate epithelium-specific Ets transcription factor, interacts with the androgen receptor and activates prostate-specific antigen gene expression. *J. Biol. Chem.* **275:** 1216–1225.

Ostvold, A.C., Holtlund, J., and Laland, S.G. 1985. A novel, highly phosphorylated protein of the high mobility group type, present in a variety of proliferating and non-proliferating mammalian cells. *Eur. J. Biochem.* **153:** 469–475.

Ostvold, A.C., Hullstein, I., and Laland, S.G. 1992. The phosphate groups of the high mobility group like protein P1 strengthens its affinity for DNA. *Biochem. Biophys. Res. Commun.* **185:** 1091–1097.

Ostvold, A.C., Norum, J.H., Mathiesen, S., Wanvik, B., Sefland, I., and Grundt, K. 2001. Molecular cloning of a mammalian nuclear phosphoprotein NUCKS, which serves as a substrate for CdK1 in vivo. *Eur. J. Biochem.* **268:** 2430–2440.

Park, I., Chung, J., Walsh, C.T., Yun, Y.D., Strominger, J.L., and Shin, J. 1995. Phosphotyrosine-independent binding of a 62-kDa protein to the src homology 2 (SH2) domain of p56(lck) and its regulation by phosphorylation of Ser-59 in the lck unique N-terminal region. *Proc. Natl. Acad. Sci.* **92:** 12338–12342.

Quakenbush, J., Liang, F., Holt, I., Pertea, G., and Upton, J. 2000. The TIGR Gene Indices: Re-construction and representation of expressed gene sequences. *Nucleic Acids Res.* **28:** 141–145.

Schuler, G.D., Boguski, M.S., Stewart, E.A., Stein, L.D., Gyapay, G., Rice, K., White, R.E., Rodriguez-Tome, P., Aggarwal, A., Bajorek, E., et al. 1996. A gene map of the human genome. *Science* **274:** 540–546.

Shin, J. 1998. P62 and the sequestosome, a novel mechanism for protein metabolism. *Arch. Pharm. Res.* **21:** 629–633.

Stekel, D.J., Git, Y., and Falciani, F. 2000. The comparison of gene expression from multiple cDNA libraries. *Genome Res.* **10:** 2055–2061.

Thaw, P., Baxter, N.J., Hounslow, A.M., Price, C., Waltho, J.P., and Craven, C.J. 2001. Structure of TCTP reveals unexpected relationship with guanine nucleotide-free chaperones. *Nat. Struct. Biol.* **8:** 701–704.

Thompson, H.G.R., Harris, J.W., Lin, F., Wold, B., and Brody, J.P. 2002. P62 over-expression in breast tumors and its regulation by Prostate-Derived Ets Factor PDEF in breast cancer cells in vitro. Preprint. Available at http://brodylab.eng.ucl.edu/~jbrody/tmp/p62.pdf.

Thompson, J., Koumari, R., Wagner, K., Barnert, S., Schleussner, C., Schrewe, H., Zimmermann, W., Muller, G., Schempp, W., and Zaninetta, D. 1990. The human pregnancy-specific glycoprotein genes are tightly linked on the long arm of chromosome 19 and are coordinately expressed. *Biochem. Biophys. Res. Commun.* **167:** 848–859.

Vadlamudi, R.K. and Shin, J. 1998. Genomic structure and promoter analysis of the p62 gene encoding a non-proteasomal multiubiquitin chain binding protein. *FEBS Lett.* **435:** 138–142.

Vadlamudi, R.K., Joung, I., Strominger, J.L., and Shin, J. 1996. P62, a Phosphotyrosine-Independent Ligand of the SH2 domain of p56(lck), belongs to a new class of ubiquitin-binding proteins. *J. Biol. Chem.* **271:** 209235–209237.

Velculescu, V.E., Madden, S.L., Zhang, L., Lash, A.E., Yu, J., Rago, C., Lal, A., Wang, C.J., Beaudry, G.A., Ciriello, K.M., et al. 1999. Analysis of human transcriptomes. *Nat. Genet.* **23:** 387–388.

Walaas, S.I., Ostvold, A.C., and Laland, S.G. 1989. Phosphorylation of P1, a high mobility group-like protein, catalyzed by casein kinase II, protein kinase C, cyclic AMP-dependent protein kinase and calcium/calmodulin-dependent protein kinase II. *FEBS Lett.* **258:** 106–108.

Walker, M.G., Volkmuth, W., Sprinzak, E., Hodgson, D., and Klingler, T. 1999. Prediction of gene function by genome-scale expression analysis: Prostate cancer-associated genes. *Genome Res.* **9:** 1198–1203.

Wheeler, D.L., Church, D.M., Lash, A.E., Leipe, D.D., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Tatusova, T.A., Wagner, L., et al. 2002. Database resources of the National Center for Biotechnology Information: 2002 update. *Nucleic Acids Res.* **30:** 13–16.

Yamada, N., Tamai, Y., Miyamoto, H., and Nozaki, M. 2000. Cloning and expression of the mouse Pse gene encoding a novel Ets family member. *Gene* **241:** 267–274.

Yawata, M., Murata, S., Tanaka, K., Ishigatsubo, Y., and Kasahara, M. 2001. Nucleotide sequence analysis of the approximately 35-kb segment containing interferon-γ-inducible mouse proteasome activator genes. *Immunogenetics* **53:** 119–129.

Zou, Z., Anisowicz, A., Hendrix, M.J.C., Thor, A., Neveu, M., Sheng, S., Rafidi, K., Seftor, E., and Sager, R. 1994. Maspin, a serpin with tumor-suppressing activity in human mammary epithelial cells. *Science* **263:** 526–529.

WEB SITE REFERENCES

ftp://ftp.ncbi.nlm.nih.gov/repository/UniGene/; Homo Sapiens UniGene data set.

Received May 10, 2002; accepted in revised form July 31, 2002.